NiCE Working Paper 09-111

August 2009

Origin and Interpretation of Internal and External Validity in Economics

Floris Heukelom

Nijmegen Center for Economics (NiCE)

Institute for Management Research

Radboud University Nijmegen

P.O. Box 9108, 6500 HK Nijmegen, The Netherlands

http://www.ru.nl/nice/workingpapers

Abstract

Validity as a key concern in experimental methodology came to the fore in American post-World War Two psychology. Following the publication of the APA's *Technical Recommendations* in 1954, validity became firmly ingrained as one of the key concepts in psychologists' methodological discourse, without however any of the many different taxonomies prevailing over the others. In the 1980s and 1990s experimental economists introduced first external and later internal validity to demonstrate the experiments' validity for economic theorizing. Subsequently, In the 2000s internal and external validity in economics became analytically defined as relating the inside world of the laboratory to the outside world of reality, and as defining a trade-off between the costs and pay-off of experimental control. Both these analytical definitions are shown to be illusory.[1]

[1] Email: F.Heukelom@fm.ru.nl.

1. Introduction

Internal and external validity are key terms in contemporary discussions on the methodology of experiments in economics. Browsing through experimental literature one easily gets the impression that these terms have always been neatly defined and uncontested tools for understanding experimentation. The present article seeks to nuance that view. The nuance unfolds along two lines. The first line, and in number of words by far the larger of the two, provides a historical account of validity, starting with its origin in statistics (section 2), through its flowering and extensive treatment in psychology (section 3), to its adoption and definition in economics (section 4). The second line consist of two critical reflections regarding the strict definition of internal and external validity in contemporary economics (section 5). The conclusion, finally, asks what lessons can be derived from these nuances (section 6).

2. The origin of validity in statistics

Validity received its most extensive in treatment in post-war American psychology, but its origins lie in interwar statistics. The two scientists involved were Ronald Fisher and Louis Leon Thurstone, the latter both statistician and psychologist. As is well known, in the 1920s Fisher almost single-handedly created modern statistical analysis for scientific experiments (e.g. Hald, 1998). With the publication of *Statistical Methods for Research Workers* in 1925 and its meticulous discussion of tools such as hypotheses testing, the Student distribution, and analysis of variances, Fisher set the standard for scientific experimentation up to the present. From the start, the statistical methods and the design of the experiments were two sides of the same coin for Fisher. Because *Statistical Methods* put the emphasis on the statistics, in 1935 the book was followed up by a sequel entitled *The Design of Experiments*, which explored in more detail the design of the experiment (Fisher, 1935, p.ii).

At the core of Fisher's research lay the question which experimental results allow for which inferences, in which both the design of the experiment and the statistical methods were crucial for the inference to be valid. The term 'valid' or 'validity' only entered two or three times in *Statistical Methods*. In the *Design of Experiment*, however, the terms entered loosely in a number of contexts. Fisher used what was a common English term to discuss such things as the design precautions needed to make the estimates based on the experimental results valid estimates, and the validity of the estimates of the errors when only a draw of the population was

taken. The term validity was a recurring concept in Fisher's two monographs, while never coming close to something like a key term or concept.

In the work of Chicago-based psychometrician Thurstone, the concept of validity became a more central element of the experimental methodology, as exemplified by Thurstone's handbook *The Reliability and Validity of Tests: derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems* (1931). Although the term did not receive an extensive treatment or clear definition, validity was, as the title indicates, one of the two guiding concepts in this 130-page student manual. Thurstone reminded the reader that "validity refers to the correlation between a test and its oriterion" (Thurstone, 1931, p.97). The example Thurstone used was a test of students' previous examination grades as a predictor for future scholarship grades. If this test performed well, in the sense that previous examination grades were a good predictor of future scholarship grades, the test had a high validity, if it did not the test had a low validity.

3. Validity in psychology

Building on the work of Fischer and Thurstone, validity as a key concern in experimental methodology came to the fore in American post-World War Two psychology. The appearance of validity as a concern in experimental practice in psychology in the early 1950s was abrupt. In Edwin Boring's classic *A History of Experimental Psychology* (1929), the issue of validity was completely absent. Also the equally voluminous and complete overview *Experimental Psychology* (1938) by Robert Woodworth and Harold Schlosberg contained no reference to validity. Moreover, validity remained absent in the 1950 second edition of Boring's history, despite the fact that the "treatment of the later period [was] greatly expanded" (Boring, 1950, p.vii). The same is true for Woodworth and Schlosberg's revised edition that appeared in 1954. Even *A Manual of Psychological Experiments* (1937), edited by Boring and with contributions from some twenty high-standing experimental psychologists of the time, in its many examples of the practicalities of conducting experiments, did not spend a single word on validity.

The principal origin of discussions of validity, then, is the American Psychological Association's (APA) *Technical Recommendations for Psychological Tests and Diagnostic Techniques* (1954), based in part on Crombach and Meehl (1955). The APA was created in 1892 and aimed at representing all American

professional psychologists (apa.org, accessed 7 May, 2009). In this capacity it actively sought to establish standards regarding various methodological aspects of psychological research. Well-known are its standards for scientific publication, first published in 1952, with precursors going back to 1929 (Kadzin, 2001, p.xix). Less successful examples include the Committee on Measurement (1909-1919).

In 1950, the then president of the APA, Joy Paul Guilford, a psychometrician working in the tradition of Thurstone, appointed a Committee on Test Standards to prepare a "statement on technical standards for evaluating tests and the contents of test manuals" (Street, 1994, p.152), which would be "an official statement of the profession" (APA Committee on Test Standards, 1952, p.461). The committee, composed of Edward S. Bordin, R.C. Challman, H.S. Conrad, Lloyd G. Humphreys, Paul E. Meehl, Donald E. Super, and its chairman Lee J. Crombach, was created to set the standard for performing tests and experiments for years to come.[2] It conducted its discussion in close cooperation with representatives of the American Educational Research Association (AERA) and the National Council on Measurements Used in Education (NCMUE). A preliminary version of its recommendations was published for examination by the psychological community in the *American Psychologist* (1952, pp.461-475). The final version was published both as a separate manual and as a special issue of the *American Psychologist* (1954, pp.201-238). About a third of this forty-page manual was devoted to validity.

The *Technical Recommendations* distinguished four types of validity: content validity, predictive validity, concurrent validity, and construct validity. Content validity was described as a claim regarding whether observations from a draw of the population justify claims about the population as a whole. Predictive validity was understood as an assessment of how well predictions based on current observations accurately reflect future observations. Concurrent validity according to the authors asked whether a new measurement instrument can distinguish between already established categories. Construct validity, finally, was an assessment of the overlap between the scientific operationalization of a higher-order term (intelligence, academic performance, unemployment, and so on) and the meaning of this higher-order term in everyday language. The *Technical Recommendations'* presentation of

[2] The psychological literature of this period alternates between using 'test' as a general term subsuming all forms of empirical investigation in psychology, including experiments, questionnaires, and developmental and clinical tests; and using 'test' and 'experiment' interchangeably when discussing the methodology of experiments.

validity was a mixed success. Almost immediately following its publication, the *Technical Recommendations'* classification of validity became contested. However, it did succeed in deeply ingraining validity as the banner under which to discuss the methodology of testing and experimenting in psychology. Moreover, through the export of psychological experimentation to the other social sciences in the postwar period, it unintentionally succeeded in ingraining the concept of validity in the social sciences generally.

Principal among the *Technical Recommendations*' contesters was Donald Campbell. In "Factors Relevant to the Validity of Experiments in Social Settings" (1957), Campbell proposed a distinction between internal and external validity.

> Validity will be evaluated in terms of two major criteria. First, and as a basic minimum, is what can be called *internal validity*: did in fact the experimental stimulus make some significant difference in this specific instance? The second criterion is that of *external validity*, *representativeness*, or *generalizability*: to what populations, settings, and variables can this effect be generalized. (Campbell, 1957, p.297, emphasis in the original)

Two aspects of Campbell's alternative distinction need emphasis. First, internal validity was given a specific meaning related to, but different from later definitions. In addition to validity as a purely statistical concept, as in the work of Fisher and Thurstone, internal validity included the issue of whether the test or experiment had produced observations not obtained by earlier tests. That is, the observations yielded by the test were internally valid only if they were statistically valid, and if they presented an insight not yet produced by earlier tests or experiments. Second, although Campbell noted that internal and external validity might sometimes be incompatible when controls for internal validity jeopardize external validity, they were not presented as opposites. A bad test or experiment could be inferior in both the internal and external validity domain, and a good test or experiment would score high on both internal and external validity.

In addition to internal and external validity, Campbell experimented with other classifications. Campbell and Donald Fiske's "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix" (1959), widely cited as the start of the double-blind research method (Brewer, 2001), argued that underneath the "major

types of validity" of the *Technical Recommendations* lay concerns for convergence and discrimination. Validation is convergent, the authors argued, when a particular test confirms observations obtained from other tests. For the validity of a new test, however, the observations in addition needed to be shown to discriminate more, or to discriminate differently as compared to previous tests. References to Campbell (1957) are absent in Campbell and Fiske (1959), so that these articles are best understood as idiosyncratic attempts at providing a taxonomy of validity in psychological experimentation.

The distinction between internal and external validity returned in Campbell and Stanley (1963), "Experimental and Quasi-Experimental Designs for Research on Teaching." In this article, external validity retained the meaning it was given in Campbell (1957). The meaning of internal validity, however, was slightly sharpened: "*Internal validity* is the basic minimum without which any experiment is uninterpretable: Did in fact the experimental treatments make a difference in this specific context?" (Campbell and Stanley, 1963, p.175, emphasis in the original). Subsequently, internal validity was divided into "eight different classes of extraneous variable," which need to be controlled in the experimental design. These eight classes included the influence of the passage of time on subjects' behavior, biases resulting from the method of selecting the subjects, and loss of responses during the experiment. External validity was divided into four classes, including the influence of previous tests on subjects' behavior, and effects of the experimental arrangements on the behavior of the subjects.

The discussions of validity in the *Technical Recommendations* and by Campbell between the early 1950s and early 1960s defined the experimental vocabulary in psychology for subsequent decades, up into the twenty-first century. They provided psychology with a range of categories of validity, some of which were sometimes considered to partly overlap, and some of which were sometimes considered to partly jeopardize each other. Validity had arisen as a main methodological concern, without however any of the existing typologies dominating the others. Let me give two contrasting examples.

In the authoritative handbook *Quasi-Experimentation, Design & Analysis Issues for Field Settings* (1979), Thomas Cook and Donald Campbell started by broadly distinguishing internal and external validity. Under the heading of internal validity they discussed what they considered a main threat to internal validity in the

case of quasi-experiments: statistical conclusion validity. A main factor of external validity, on the other hand, was construct validity. Thus, internal and external validity functioned as the general terms under which other forms of validity, but by no means all existing sub-classes found their place. Similarly, in the introductory textbook *The Psychologist as Detective, An Introduction to Conducting Research in Psychology* (1997), Randolph Smith and Stephen Davis spent an entire chapter, or about thirty-five of the book's 450 pages discussing validity. They framed their discussion conveniently in internal and external validity, without however any recourse to one of the other types of validity.

In sharp contrast to these authors, the 2000 eight-tome version of the APA's *Encyclopedia of Psychology* (Kazdin, 2000), went in a completely different direction. It contained two entries for validity, "Validity," and "Construct Validity," signaling a predominant position of construct validity among the different categories. Construct validity in its entry was discussed without reference to other types of validity (Han, 2001). The more general entry on validity, on the other hand, distinguished along construct validity, four types of validity: content, face, criterion, and consequential validity. In both entries there was not a single reference to internal or external validity.

It is not the purpose of this article to spell out conflicting schools of thought in psychological methodology. Instead, this section's brief historical account merely wants to reveal three characteristics of the concept of validity in psychology. First, following Fisher and Thurstone's tentative introductions of the term, and in particular since the publication of the APA's *Technical Recommendations* in 1954, validity became firmly ingrained as one of the key concepts in psychologists' methodological discourse. Second, over its eight decades of existence, validity evolved from a purely statistical concept, to a broader concept of control in psychological empirical investigation. Third, despite, or perhaps because of its central place in the psychologists' methodological discourse, many different definitions and categorizations abounded, without any of the taxonomies prevailing over the others.

4. The dissemination of validity into economics

Once established as a major aspect of experimental practice, the use of validity disseminated from psychology into other social sciences, including economics and sociology. Because validity was connected to testing and experimentation, it is no surprise that the dissemination of validity into economics occurred through

economists' use of psychologists' experimental method.[3] Thus, the history of validity in economics starts rise of experimental economics.

4.1 Validity in experimental economics

It is well known that Vernon Smith, the founding father of experimental economics, developed his experimental method for economics in an atmosphere of interdisciplinary research in the 1950s (Dimand, 2005, Weintraub, 1992, Lee, 2004). In particular, Smith's collaboration with social psychologist Sidney Siegel is recognized as vital for the development of Smith's application of the experimental method of the psychologists to economic research questions (Innocenti, 2008, Lee, 2008, Smith, 1992). However, among others because of Siegel's death in 1961, Smith was alone in conducting his economic experiments until the mid-1970s, when he was joined by a younger generation of economists, including Charles Plott, Dan Friedman, and many others.

Discussions of the proper method of conducting experiments in economics are a recurring theme in Smith's writings of the 1960s and in the experimental economics literature more generally from the mid-1970s onwards. Yet, these discussions were not put in terms of the validity framework that was developing in psychology around the same time. On the contrary, in the single methodological article Smith wrote in the 1960s, the problem of inferences based on experiments is a few times put in terms of validity, without however any reference to the psychologists, and despite the fact that in the same article Smith explicitly cites some social psychologists as important for developing the experimental method for economics. Instead, Smith took a position which is best described as a combination of set-theory and Bayesian statistics, and which derived its inspiration mainly from Savage (1962). In this approach, the overlap of the sets Nature, Model, and Experimental Results "constitutes the 'validity' of the experiment" (Rice and Smith, 1964, p. 240), the statistics of which are an input for a Bayesian updating process calculating the scientist's post-experiment beliefs.

Also in Smith's well-known methodological article of the 1970s, "Experimental Economics: Induced Value Theory" (Smith, 1976), no reference was made to the validity framework of the psychologists. Smith talked about the

[3] In this paper I focus on laboratory experiments in experimental and behavioral economics. Recent developments such as experimenting in financial accounting (e.g. Libby, Bloomfield, and Nelson, 2002).

experiments as testing the "validity" of economic theories (Smith, 1976, p.274), invokes the concept of parallelism ("As far as we can tell, the same physical laws prevail everywhere" (Harlow Shapley, 1964, quoted in Smith, 1976, p.274)), and advanced control as "the essence of experimental methodology" (Smith, 1976, p.275). But validity as understood by the psychologists did appear not even in a footnote.

The reason that Smith and other experimental economists in the 1970s, such as Plott, were reluctant to adopt validity as understood by the psychologists, was that the psychologists' way of dealing with validity risked creating a division between an inside world of the laboratory and an outside "real" world of the economy and its actors. Such an interpretation, they argued, would be directly against experimental economists' conception of experiments in economics because,

> The relevance of experimental methods rests on the proposition that laboratory markets are 'real' markets in the sense that principles of economics apply there as well as elsewhere. Real people pursue real profits within the context of real rules. The simplicity of laboratory markets in comparison with naturally occurring markets must not be confused with questions about their reality as markets. (Plott, 1982, p. 1520)

For this reason, Plott (1982) was equally dismayed about the term "artificial" as a description of the laboratory environment. Plott argued that a concern with the artificiality of experiments is a straw man, and that in any case it would be an argument directed at experimentation in general, not just at experiments in economics.[4] Finally, also the 560-page leading textbook in experimental economics, Davis and Holt *Experimental Economics* (1993), was entirely devoid of references to psychologists' validity.

Yet, in the late 1980s and early 1990s the psychological categories nevertheless slowly started to creep into the methodological discussion of the experimental economists. The first article to invoke one of the psychologists' notions of validity was Brookshire, Coursey and Schulze, "The External Validity of Experimental Economics Techniques: Analysis of Demand Behavior" (1987). The

[4] Similar arguments such as these in Plott (1982), can be found in Smith (1976, 1982), and Wilde (1980). The most extensive presentation and refutation of artificiality and the associated real world argument can be found in Starmer (1999).

focus of the authors were the now well-known twin allegations of experimental economic results not translating to "real world settings," and how to compare behavior of student-subjects to the behavior of "'actual' buyers, sellers, or traders" (p.289). In contrast to earlier replies to these allegations, as indicated above, the authors summarized them as focusing upon "the *external validity*, or lack of validity, associated with experimental economic techniques" (p.289, emphasis in the original), which was then equated with the issue of the parallelism between the two settings, as defined by Smith and others. Subsequently, the authors set out to invalidate these allegations on empirical grounds. That is, they compared subjects' behavior in experimental markets with agents in 'real' markets and concluded that the observed behavior in the two settings was similar enough to conclude that the experimental method in economics did not pose a problem of parallelism or external validity.

It is first of all important to note that the authors adopted external validity without its counterpart internal validity, thus defying the idea that the two always come together. Second, the issue of external validity was introduced in experimental economics not as a fundamental philosophical problem of experimentation, but as a purely empirical question. This continues to be the understanding of (external) validity of many experimental economists up to this day. For instance, in an email to the author, Smith emphasized that he considered validity, or parallelism, "100% an empirical issue," an issue that "needs to be addressed as such if wheel spinning is to be avoided" (email Smith to author, 1 July 2009).

Following this first experimental economic article employing external validity, also internal validity was introduced. For instance, in Friedman and Sunder *Experimental Methods, A Primer for Economists* (1994), economists interested in conducting experiments were occasionally warned that such-and-so method may undermine or threaten the internal validity of the experiment, or that some other procedure risked weakening the external validity of the experiment, in which external validity continued to be equated with parallelism. As a result, internal and external validity gradually became household concepts in experimental methodology in economics.

Nevertheless, many experimental economists continued to have some reservations with the new methodological concepts from psychology. Vernon Smith tried to play down the importance of the two validities, emphasizing, as said, that it was merely an empirical question that could be tested, and continuing to equate

validity with parallelism. In an email to the author Dan Friedman noted that he has never been "especially enthusiastic about [..] 'internal' and 'external' validity" (email Freidman to author, 2 July 2009). Similar remarks have been made by other well-known experimentalists (email Glenn Harrison to author, 2 July 2009, email John List to author, 3 July 2009).

The question thus arises why these experimental economists adopted the internal and external validity concepts of the psychologists if they apparently did not like them all that much. The most plausible answer, as illustrated by Brookshire, Coursey and Schulze's article discussed above, is that the experimental economists of the late 1980s and early 1990s felt compelled to use internal and external validity in order to establish experimental economics as a viable economic sub-discipline. As experimental economics was growing but its position anything but secured, external and later internal validity were introduced to convince a skeptical audience of economists who did not believe in the 'reality' of experiments (Starmer, 1999, email Chris Starmer to author, 1 July 2009, email Friedman to author, 2 July 2009). It is hence somewhat ironic that the use of internal and external validity subsequently ingrained in economics the conception of an inside world of the laboratory versus an outside world of reality.

4.2 Behavioral economics

The introduction of external and internal validity by experimental economists in the late 1980s and early 1990s, coincided with the emergence of another economic sub-discipline which employs experiments (Heukelom, 2009). Despite the fact that during the 1990s and early 2000s behavioral economics and experimental economics came to oppose one another on a number of issues (Heukelom, forthcoming), behavioral economists adopted many of the experimental techniques and language developed by the experimental economists, including an aversion towards the use of deception, the use of monetary rewards, and the use of internal and external validity.[5] However, behavioral economists dropped the link between external validity and parallelism and preferred to refer to internal and external validity as a "psychological distinction." As in experimental economics, validity was not extensively discussed in behavioral economics, as compared to the extensive discussions in psychology. When it was

[5] Arguable, the distinction between experimental and behavioral economics started to dissolve around the mid-2000.

discussed, however, it was framed in terms of internal validity and external validity, in which internal validity, roughly and without much discussion referred to the validity of the inferences drawn on the basis of the experimental observations, and external validity referred to the generalizability of the observations and inferences.

In a wonderful historical twist, Loewenstein (1999), the most extensive and explicit behavioral economics discussion of validity in the 1990s used the "psychological distinction" of internal and external validity to criticize experimental practice of experimental economists. "Experimental Economics from the vantage-point of Behavioural Economics," Loewenstein (1999), positioned behavioral economics explicitly in opposition to experimental economics. Under the heading of external validity, Loewenstein saw four problems with experimental economics. First, experimental economics put great emphasis on the use of auctions in its experiments. As people in reality hardly ever find themselves in an auction situation, it is doubtful that these experiments can tell us very much about economic behavior in the real world. Second, Loewenstein disagreed with experimental economists' use of repetition in what he called the Ground Hog Day argument, following Camerer (1996). In reality, Loewenstein argued, people never make the exact same decision forty times in a row. Real world behavior is much more like the first few rounds of an experiment than the last two or three rounds. Third, Loewenstein criticized experimental economists for their tendency to reduce real-world content to the absolute minimum possible. Apart from the fact that a context-free experiment is an illusion, Loewenstein argued it also greatly reduces the external validity of the experiments.[6] Instead, economists should, just as Loewenstein himself, make the experimental situation as congruent with reality as possible; hence make the experiment "context-rich." Fourth, according to Loewenstein experimental economists wrongly assumed that monetary rewards result in strict control over incentives. With monetary incentives, subjects are also likely to be driven by other motives than profit maximization, he argued. Finally, one problem concerning internal validity that Loewenstein observed was that experimental economists had been far too careless in not using randomization and in comparing the experimental results that had been obtained under different circumstances.

[6] Loewenstein both uses 'context' and 'content.'

Apart from the question of whether these criticisms are justified, Loewenstein's discussion illustrates that by the late 1990s internal and external validity had emerged as household concepts for experimental methodology in both experimental and behavioral economics. They were not given extensive treatment or definition, but definitively had emerged as two contrasting concepts in which terms experimentalists in economics thought about their experiments.

4.3. The analytical definition of internal and external validity in economics

Thus, in the 1980s and 1990s external and internal validity were loosely introduced in the methodology of experiments in economics. Although the economists who conducted and discussed seemed to be comfortable with the meaning of both terms and with the relationship between the two, internal and external were never given extensive discussion or definition. This changed in the late 1990s and early 2000s when philosopher of science *cum* experimental economist Francesco Guala applied his analytical skills to the issue of validity in economics (Guala, 1999, 2003, 2005, Guala and Mittone, 2005).

Guala first of all drew a sharp line between an inside world of the experiment and an outside, real world, thus implicitly arguing against the position of the experimental economists of the 1960s-1980s, but in line with discussion the approach of the psychological community. Experimentalists draw "inferences *within* the experiment," Guala argued, in which experiments are to be understood as "very special settings, which are rarely if ever instantiated in the 'real world' outside the laboratory" (Guala, 2005, p.141, emphasis in the original). Internal validity, according to Guala, was about the validity of the inferences in the inside world of the laboratory, external validity was about the relation of these inferences to the real outside world. Internal validity "is achieved when some particular aspect of a laboratory system [..] has been properly understood by the experimenter" (Guala, 2005, p.142). Experiment E is internally valid when variation in Y is known to be caused by X. It is in addition externally valid if "X causes Y not only in E, but also in F,G,H, etc" (Guala, 2005, p.142). As a result, claims about internal validity should be understood as chronologically and epistemically antecedent to claims about external validly. Second, Guala posited a trade-off between internal and external validity:

> The stronger an experimental design is with respect to one validity issue, the weaker it is likely to be with respect to the other. The more artificial the environment, the better for internal validity; the less artificial, the better for external purposes. (Guala, 2005, p.144)

Guala's work has been taken up by a number of experimentalists (e.g. Schram, 2005, Bardsley at al., 2009). For instance, Schram (2005) "Artificiality: The Tension Between Internal and External Validity in Economic Experiments" constructs a similar framework as Guala, and relied for its terminology furthermore on Loewenstein (1999). Like Guala, Schram explicitly posited a trade-off between internal and external validity, and retraces it directly to psychology.

> There is an obvious tension between [internal and external validity]. Where internal validity often requires abstraction and simplification to make the research more tractable, these concessions are made at the cost of decreasing external validity. Loewenstein (1999) points out that while this tension is a starting point in learning research methods in psychology, the discussion is often completely neglected by economists. (Schram, 2005, p.226)

In addition, Schram like Guala invoked a distinction between the laboratory world of the experiment and the real world outside the experiment. A key issue in economic methodology of experimentation, according to Schram, is the issue of the "artificiality of the laboratory situation," in which "the [artificiality] question is whether the stylized form of experimental institutions allow for conclusions pertaining to the 'real world'" (Schram, 2005, p.226). As said, in the experimental economic literature the outside versus inside dichotomy is often linked to this notion of "artificiality" (e.g. Starmer, 1999, Schram, 2005, Bardsley, 2005), in which artificiality refers to the alleged artificial world within the laboratory that is different, and therefore perhaps not comparable, to the "real," or "natural" world outside the laboratory. In other words, the problem of external validity is understood as the problem of artificiality.

With the analytical treatment by Guala and others, internal and external validity in economics came to the center of a logic of experimentation in economics. Where experimental economists of the 1980s and early 1990s such as Vernon Smith

conceived of validity as a purely empirical question of the comparability between experimental outcomes in different settings, in the 2000s validity became the way to logically connect the experimental setup to the experimental results, and to logically connect the experimental results derived in the 'inside' world of the experiment to the 'outside' world of reality. At the end of the first decade of the twenty-first century, with the distinction between experimental and behavioral economics gradually dissolving and with other experimental and empirical procedures emerging, the analytical take on validity in economics seems to have the upper hand. Despite a number of experimental economists who resist or who are uncomfortable with the logic of experimentation in terms of internal and external validity, it is that logic that provides the dominant way of thinking about experimentation in economics.

Three aspects of the dissemination of validity from, first, statistics to psychology and, second, from psychology into economics are noteworthy. First, when validity emerged in the work of Fischer and Thurstone it was a concept without any further division. In post-war American psychology, however, validity was quickly divided in first four, and subsequently many more sub-categories, between which the relationship was not always clear. From this wide variety of categories economics adopted only two, first external and later internal validity. Despite its prominence in post-war statistical and experimental reasoning, validity has been specified and employed in different ways in the different disciplines. Second, the definition of internal and external validity in economics became stricter than it had been in psychology. Although psychologists recognized that elements of internal and external validity could jeopardize one another, it was only when they migrated to economics that internal and external validity were understood as opposing and excluding categories. Third, the treatment of internal and external validity in economics in recent years has provided an argument in the distinction between laboratory experiments and field experiments (e.g. Harrison and List, 2004, Carpenter, Harrison and List, 2005). Basically, the reasoning is that laboratory experiments allow for a relatively large amount of control and therefore provide a high degree of internal validity. The costs of this high internal validity, however, is that laboratory experiments yield a relatively low external validity. Experimental practice in economics therefore leaves room for a method which off-sets the low external validity of laboratory experiments. This is were field experiments come in, which lack the

high internal validity of laboratory experiments, but therefore offer a high degree of external validity.

5. Two critical reflections

The previous three sections have provided a history of validity in statistics, psychology, and economics. I would like to use this last section to advance two critical reflections regarding economists' use of validity.

5.1 Internal versus external wrongly suggests an inside versus an outside

It is probably difficult to altogether dispense with a distinction between an 'inside' world of the laboratory and an 'outside' world of reality. The intuition behind such a distinction is that in an experimental setting, i.e. in a laboratory, the scientist controls some factors of the world and thereby creates an artificial world inside the laboratory. When such a distinction is used loosely and without any specific, logical interpretation attached to it, there does not seem to be any harm in employing it. However, when it becomes part of a strict analytical account of experimentation such as it receives in the work of Guala and its followers, it does become problematic. The reason is that a distinction between an 'inside' world of the laboratory and an 'outside' world of reality confuses the physical operation of conducting an experiment with its ontological status. True, experiments often are physically conducted 'inside' a laboratory. The scientist puts on her gloves and glasses, passes through an airlock and enters the laboratory. In economics and psychology similarly the scientist often physically enters the laboratory where the experiment is conducted. However, that does not mean that by doing so she enters a different, inside, or non-real world, as opposed to the real world outside. It simply means that she enters a place in the world specifically manipulated and controlled to conduct experiments. In a laboratory the scientist is completely in our real, material world.

To put the matter differently, what we call laboratories are sites best equipped to obtain the desired experimental manipulation and control. But if the mere fact of physical control and manipulation of certain aspects of the world would mean that we are in a different, or inside world, then every time an organism tries to manipulate or control some aspects of the world, a new different inside world would be created. It would mean that when I bake my favorite apple pie on a Saturday afternoon, and in the process manipulate and control certain aspects of the world (mixing the

ingredients, setting the oven, and so on), I would be creating a new inside world. Obviously, it is nonsense to think of my backing the apple pie in such a way. Man's attempt to control and manipulate aspects of the world creates an artifact – something man-made that did not yet exist – and it may create something that cannot exist on earth outside the laboratory or my kitchen. But it always remains a real part of our material world.

5.2 The trade-off between internal and external validity is an illusion

Both Loewenstein (1999), Guala (Guala, 1999, 2003, 2005, Guala and Mittone, 2005) and Schram (2005) suggest a trade-off between internal and external validity, in which more of the one by definition implies less of the other. Such a trade-off is an illusion, and indicates an economic mind-set in which control comes at a cost. Why such a trade-off is an illusion is best illustrated by means of a counter example. Suppose I want to conduct an experiment on altruism. I apply for a grant, which I receive, and I set up a dictator game experiment with a variety of treatments. Now, I suspect there may be some variation in altruistic behavior in different subgroups of the human population. There may be a difference between men and women, between Europeans and Americans, between rich and poor, between older and younger people, between students of economics and students of other disciplines, and so on. However, because of institutional and monetary constraints I am forced to conduct my experiment with first year students of economics, a majority of which are men. Obviously, this reduces the external validity of my experiment. Inferences about altruistic behavior in human beings on the basis of my experiment are less valid as compared to a situation in which I would have used a more representative sample of the human population. But in no way has this relatively low external validity of my experiment *by definition* been off-set by a higher internal validity. It may of course be true that in this particular case the internal validity has increased, for instance because my pool of to-be economists can be assumed to more readily understand the instructions or because they can be assumed to respond similarly to the incentive structure I have designed. But these possible increases in internal validity have not happened *because* of the lower external validity. There is no causal connection between the two. It is equally possible that my selected pool of economic students leads to a lower internal validity than would have been the case with a more representative draw of the population, for instance because the students were more

than the average affected by the late Friday afternoon time of my experiment, or by the good-looking female PhD student who runs the experiment. In short, there is no necessary trade-off between the external and internal validity of my experiment. In the worst case, I will have to conclude in the end that my experiment has scored poorly on both the external and the internal validity. In a less dramatic case I might be able to conclude that despite the not so perfect external validity, my experiment was flawless on the internal validity concerns. As stressed by Campbell when he first introduced the internal and external validity concepts, it may be a challenge to achieve both a high internal and external validity. However, the idea that the two categories are opposites and that they can be traded off is simply wrong.

6. Conclusion

The purpose of this article, as set out in the introduction, has been to nuance the understanding of internal and external validity in contemporary economics by providing a historical account of the origins and development of validity, and by providing two critical reflections on our understanding of internal and external validity in economics. What lessons can be drawn from these nuances?

It is easiest to start by what should not be done. First of all, we economists should try to avoid speaking in terms of an inside world of the laboratory versus an outside world of reality – or language along those lines. Although perhaps unproblematic when used in a loose, everyday manner of discussing experiments, strictly speaking it does not make sense. Second, economists should recognize that the idea of a necessary trade-off between internal and external validity is simply wrong, and that David Campbell was right when he introduced the concepts. Attempts to increase the one may jeopardize the other, but there is no necessary trade-off. However, that does not mean, thirdly, that economists should simply divert to psychologists' use of the term. As fifty years of extensive discussion in psychology has not produced anything approaching a consensus regarding validity, there is little to gain from adopting methodological discussion in psychology lock stock and barrel.

What we should do when we face some experimental claim regarding the economic world or regarding one of our theories about that economics world, is what we always we when we see a model, econometric analysis, questionnaire result, or opinion by one of our profession's members: ask for proof and ask for evidence. Despite the beauty of logical structures that relate theories, to models, to experiments,

and so forth, the only thing we need to ask ourselves in the end is whether we are convinced. In the end the validity of experimental results is always an empirical claim that either convinces us or that does convince us.

References

A Joint Committee of the American Psychological Association, A. E. R. A., and National Council on Measurements used in Education (1954). "Technical Recommendations for Psychological Tests and Diagnostic Technics." Psychological Bulletin **51**(2): Supplement.

Bardsley, N. (2005). "Experimental economics and the artificiality of alteration." Journal of Economic Methodology **12**(2): 239-252.

Bardsley, N., R. Cubitt, et al. (2009). Experimental Economics: Rethinking the Rules. Princeton, Princeton University Press.

Biagioli, M., Ed. (1999). The Science Studies reader. New York, Routledge.

Boring, E. C. (1929). A History of Experimental Psychology. New York, The Century Co.

Boring, E. C., Ed. (1937). A Manual of Psychological Experiments New York, John Wiley.

Boring, E. C. (1950). A History of Experimental Psychology, 2nd edition. New York, Appleton-Century-Crofts.

Brewer, M. B. (2001). Donald Campbell. Encyclopedia of Psychology. A. E. Kadzin. Washington D.C., American Psychological Association**:** III: 3-5.

Camerer, C. (1996). Rules for Experimenting in Psychology and Economics, and why they differ. Experimental Studies of Strategic Interaction: Essays in Honor of Reinhard Selten. W. Albers, W. Guth and E. Van Damme. Berlin, Springer-Verlag.

Campbell, D. T. (1957). "Factors Relevant to the Validity of Experiments in Social Settings." Psychological bulletin **54**: 297.

Campbell, D. T. and D. Fiske (1959). "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." Psychological bulletin **56**: 81.

Campbell, D. T. and J. C. Stanley (1963). Experimental and quasi-experimental designs for research. Handbook of research on teaching N. L. Gage. Boston, Houghton Mifflin.

Carpenter, J. P., G. W. Harrison, et al. (2005). Field Experiments in Economics: An Introduction. Research in Experimental Economics Volume 10. J. P. Carpenter, G. W. Harrison and J. A. List. Amsterdam, Elsevier**:** 1-15.

Cook, T. D. and D. T. Campbell (1979). Quasi-experimentation : design & analysis issues for field settings. Boston, Houghton Mifflin Co.

Crombach, L. J. and P. E. Meehl (1955). "Construct Validity in Psychological Tests." Psychological Bulletin **52**: 281.

Davis, D. D. and C. A. Holt (1993). Experimental Economics. Princeton, Princeton University Press.

Dimand, R. W. (2005). Experimental economic games: the early years. The Experiment in the History of Economics. P. Fontaine and R. Leonard. New York, Routledge**:** 5-24.

Fisher, R. A. (1925). Statistical Methods for Research Workers. Edinburgh, Oliver & Boyd.

Fisher, R. A. (1935). The Design of Experiments. Endiburgh, Oliver & Boyd.

Friedman, D. and S. Sunder (1994). Experimental Methods: A Primer for Economists. Cambridge, Cambridge University Press.

Guala, F. (1999). "The Problem of External Validity (or 'Parallelism') in Experimental Economics." Social Science Information **38**: 555-573.

Guala, F. (2003). "Experimental Localism and External Validity." Philosophy of Science **70**: 1195-1205.

Guala, F. (2005). The Methodology of Experimental Economics. Cambridge, Cambridge University Press.

Guala, F. and L. Mittone (2005). "Experiments in Economics: External Validity and the Robustness of Phenomena." Journal of Economic Methodology **12**(4): 495-515.

Han, K. (2001). Construct validity. Encyclopedia of Psychology. A. E. Kadzin. Washington D.C., American Psychological Association**:** III: 281-283.

Harrison, G. and J. A. List (2004). "Field Experiments." Journal of Economic Literature **27**: 1013-1059.

Heukelom, F. (2009). Kahneman and Tversky and the Making of Behavioral Economics. Economics. Amsterdam, University of Amsterdam. **PhD:** 170.

Innocenti, A. (2008). "How Can a Psychologist Inform Economics? The strange case of Sidney Siegel." DEPEID Working papers **8/2008**.

Kadzin, A. E., Ed. (2001). Encyclopedia of psychology. Washington D.C., American Psychological Association.

Kahneman, D. and A. Tversky (1979). "Prospect Theory: An Analysis of Decision under Risk." Econometrica **47**: 313-327.

Latour, B. and S. Woolgar (1979). Laboratory life : the social construction of the scientific facts. Beverly Hills, Sage Publications.

Lee, K. S. (2004). Rationality, Minds, and Machines in the Laboratory: A Thematic History of Vernon Smith's Experimental Economics. Notre Dame, University of Notre Dame. **PhD:** 300.

Levitt, S. D. and J. A. List (2009). "Field experiments in economics: The past, the present, and the future " European Economic Review **53**(1): 1-18.

Libby, R., R. Bloomsfield, et al. (2002). "Experimental research in financial accounting." Accounting, Organization and Society **27**: 775-810.

Loewenstein, G. (1999). "Experimental Economics from the vantage-point of Behavioural Economics." The Economic Journal **109**: F25-F34.

Morgan, M. S. (2003). Experiments without Material Intervention: Model Experiments, Virtual Experiments, and Virtually Experiments. The Philosophy of Scientific Experimentation. H. Radder. Pittsburgh, University of Pittsburgh Press**:** 216-235.

Plott, C. R. (1982). The Application of Laboratory Experimental Methods to Public Choice. Collective Decision Making: Applications from Public Choice Theory. C. S. Russell. Baltimore, Johns Hopkins University Press.

Rice, D. B. and V. L. Smith (1964). "Nature, the Experimental Laboratory, and the Credibility of Hypotheses." Behavioral Science **9**(3): 239-246.

Schram, A. (2005). "Artificiality: The Tension Between Internal and External Validity in Economic Experiments." Journal of Economic Methodology **12**(2): 225-237.

Smith, R. A. and S. F. Davis (1997). The Psychologist as Detective : An Introduction to Conducting Research in Psychology. Upper Saddle River, N.J, Prentice Hall.

Smith, V. L. (1976). "Experimental Economics: Induced Value Theory." American Economic Review **66**: 274-279.

Smith, V. L. (1982). "Microeconomic Systems as an Experimental Science." American Economic Review **72**: 923-955.

Smith, V. L. (1992). Game Theory and Experimental Economics: Beginnings and Early Influences. Towards a History of Game Theory. E. R. Weintraub. London, Duke University Press**:** 241-282.

STANDARDS, A. C. O. T. (1952). " Technical Recommendations for Psychological Tests and Diagnostic Techniques: Preliminary Proposal " The American Psychologist **1952**: 461.

Starmer, C. (1999). "Experiments in Economics: Shuold we Trust the Dismal Scientists in White Coats?" Journal of Economic Methodology **6**(1): 1-30.

Thurstone, L. L. (1931). The Reliability and Validity of Tests: derivation and interpretation of fundamental formulae concerned with reliability and validity of tests and illustrative problems Ann Arbor, Edwards Bros.

Tversky, A. and D. Kahneman (1974). "Judgment under Uncertainty: Heuristics and Biases." Science **185**: 1124-1131.

Weintraub, E. R., Ed. (1992). Toward a History of Game Theory. London, Duke University Press.

Wilde, L. (1980). On the use of Laboratory Experiments in Economics. The Philosophy of Economics. J. Pitt. Dordrecht, Reidel.

Woodworth, R. C. and H. Schlosberg (1938). Experimental Psychology. New York, Holt, Rinehart and Winston.